Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 31 (2014) 398 – 405

Procedia Computer Science

2nd International Conference on Information Technology and Quantitative Management, ITQM 2014

Text Categorization Based on Clustering Feature Selection

Xiaofei Zhou[a], Yue Hu[a], Li Guo[a]*

[a]Institue of Information Engineering, Chinese Academy of Science, Beijing, 100095, China

Abstract

In this paper, we discuss a text categorization method based on k-means clustering feature selection. K-means is classical algorithm for data clustering in text mining, but it is seldom used for feature selection. For text data, the words that can express correct semantic in a class are usually good features. We use k-means method to capture several cluster centroids for each class, and then choose the high frequency words in centroids as the text features for categorization. The words extracted by k-means not only can represent each class clustering well, but also own high quality for semantic expression. On three normal text databases, classifiers based on our feature selection method exhibit better performances than original classifiers for text categorization.

Keywords: Feature selection, text categorization, k-means

1. Introduction

In text analysis, a document is a feature vector of word-weights[1] The dimensionality of feature vector is often very large, but actually, the useful features in a class corpus are often limited into a small subset. Feature selection is an effective way to reduce the dimensionality and find important words for text expression. Statistical feature selection methods usually adopt feature search and feature evaluation strategies to remove the redundant or unimportant features[2]. Feature search method attempts to find an optimal subset of features that will provide greater class decision, such as SBS (sequential backward selection) and SFS(sequential forward selection)[3]. However, feature search method may be not the best approach when there are interacting features

* Corresponding author. Tel.: +86-010-82546701; fax:+86-010-82546701.
E-mail address: guoli@iie.ac.cn

doi:10.1016/j.procs.2014.05.283

in the dataset. Feature evaluation method usually considers an estimation computing for each feature or a feature set, and then selects the most important features. But it is difficult to choose an effective criterion for feature evaluation in clustering[4]. In past few years, many evaluation methods to choose observed words with good statistical properties used in feature selection, such as DF (Document Frequency), DIA, Chi-square, IG (Information Gain), RS (relevancy score), OR (Odds Ratio), GSS coefficient, etc[5]. For classification problem, the features with good class-representation are much significant, but current feature evaluation methods usually ignore it. In this paper, we will utilize k-means clustering method to collect the features related to the corresponding class, which avoid direct feature search and feature evaluation for each features. The cluster centroids from each class can express text category characters, and in which the corresponding items with high weights are very relevant to the class, thus we choose such features as text representation. In this paper, we discuss two metrics, cosine distance and Euclidean distance in k-means process for features collections, and then separately conduct three classifiers, k-NN, NC and SVM based on the chosen features for text categorization.

The remaining of this paper is organized as follows. In section 2, we introduce k-means with cosine distance and Euclidean distance, and then give our feature selection method. In section 3, we present some comparative experimental results on several text corpuses. At last, the conclusion and acknowledgement are given in the end.

2. k-means Feature Selection

2.1. K-means

K-means is one of the simplest clustering algorithms to group data, which aims to partition the samples into k sets with minimizing cluster error. In k-means there are three main steps, first selecting k initial cluster centroids, second assigning each sample to the nearest centroid, and final updating the centroids by the means for each cluster. We briefly give the process of k-means:

(1) Initial cluster centroids $(\mathbf{m}_1, \mathbf{m}_2, .. \mathbf{m}_k)$ are randomly selected from given samples $(\mathbf{x}_1, \mathbf{x}_2, .. \mathbf{x}_n)$.
(2) The similarities between each sample and all centroids are computed, and then each sample is assigned to the nearest centroid.
(3) The means of samples in each cluster are calculated as the new cluster centriods.

The step (2) and (3) are repeated until the final stable clustering results are obtained.

For text data, the similarity between a sample and a centroid in k-means usually adopts Euclidean distance and Cosine distance. In the following, we give the two distances.

- Euclidean distance

$$D(x_i, m_j) = \sqrt{\sum_{l=1}^{v} (x_{il} - m_{jl})^2}, \quad i = 1,\ldots,N; j = 1,\ldots,k \tag{1}$$

- Cosine distance

$$D(x_i, m_j) = \frac{\mathbf{x}_i^T \mathbf{m}_j}{|\mathbf{x}_i| \|\mathbf{m}_j|}, \quad i = 1,\ldots,N; j = 1,\ldots,k \tag{2}$$

$(\mathbf{x}_1, \mathbf{x}_2, .. \mathbf{x}_n)$ are samples, and $(\mathbf{m}_1, \mathbf{m}_2, .. \mathbf{m}_k)$ are the clustering centroids. The distance between sample and centroid adopted in k-means directly affects the clustering results, and the final centroids will have the minimal means of distances.

2.2. Feature selection

In this paper, we present to use k-means method to collect features from cluster centroids of each class. We choose the features with larger weight values in each centroid, and then take all the selected features for text categorization. The process of our feature selection is shown in Fig.1.

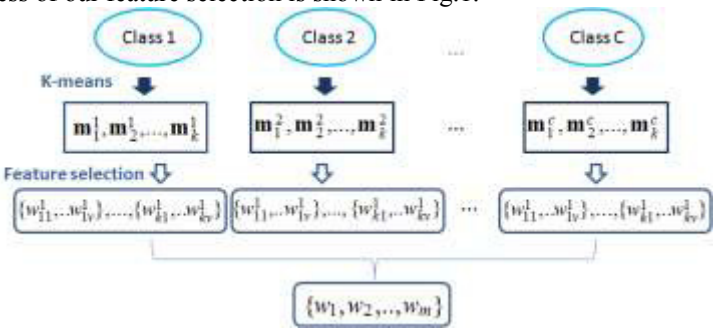

Fig. 1. Process of text feature selection

In the following, we give the steps of our feature selection algorithm.

- **Step1:** We capture k centroids for each class. The distance between sample and centroid in k-means uses Euclidean distance and cosine distance.
- **Step2:** For each centroid, we rank its features according to descend order of weight values, and select the largest v features.
- **Step3:** At last, for all the selected features, we remove the repeated ones, and get the final feature set for text expression.

3. Experiments

In this paper, based on k-means feature selection we will discuss three classifiers, nearest centroid(NC) method, k nearest neighbor(k-NN) method and SVM method for text categorization. The experiments are conducted on four text corpus, DBWorld[6]. Transcripts (Subset of Reuters Transcribed)[7], WebKB(World Wide Knowledge Base)[8]. Farm-ads[9]. We split each corpus into two parts as training and test set. The class amount, dimensionality, and data scale are given in Table 1.

Table 1. Experimental Dataset

Datasets	Class	Dimension	Train-N	Train-N
DBWorld	4	4703	128	128
Transcripts	10	6327	100	100
WebKB	4	7287	2084	2084
Farm-ads	2	54877	2071	2072

On training set, we conduct our algorithm to select features for text expression, and then represent all the samples by the selected features. After the feature selection, we run three classifiers, NC, k-NN and SVM. We compared the methods based on the selected features with original classifiers. The normal accuracies of text

categorization, macro-F score and micro-F score, and the running time of classification are tested. In Table 2, we give the comparison results.

Table 2. Experimental Results on DBWorld dataset

Methods	Macro-F (%)	Micro-F (%)	Time (s)
NC	83.73	84.38	0.658
Euclidean KMF + NC	**87.41**	**87.50**	0.016
Cosine KMF+ NC	86.47	86.72	0.031
k-NN(k=5)	43.38	54.69	8.467
Euclidean KMF + kNN	67.49	68.75	0.719
Cosine KMF + kNN	**83.24**	**83.59**	0.828
SVM	80.25	81.25	9.565
Euclidean KMF + SVM	88.90	89.06	2.000
Cosine KMF + SVM	**91.27**	**91.41**	1.797

Table 3. Experimental Results on Transcripts dataset

Methods	Macro-F (%)	Micro-F (%)	Time (s)
NC	44.65	46.00	0.978
Euclidean KMF + NC	47.30	48.00	0.203
Cosine KMF+ NC	**48.49**	**49.00**	0.406
k-NN(k=5)	25.50	30.00	6.806
Euclidean KMF + kNN	38.48	39.00	0.938
Cosine KMF + kNN	**41.16**	**43.00**	1.391
SVM	60.57	61.00	8.710
Euclidean KMF + SVM	**64.92**	**65.00**	3.828
Cosine KMF + SVM	62.38	62.00	4.047

Table 4. Experimental Results on WebKB dataset

Methods	Macro-F (%)	Micro-F (%)	Time (s)
NC	70.04	70.54	13.987
Euclidean KMF + NC	**70.38**	**70.68**	1.641
Cosine KMF+ NC	70.31	70.63	2.047
k-NN(k=5)	56.37	64.97	1491.091
Euclidean KMF + kNN	**69.32**	**73.85**	347.656
Cosine KMF + kNN	67.84	73.13	586.375
SVM	**89.07**	**89.83**	1161.272
Euclidean KMF + SVM	87.32	88.20	459.016
Cosine KMF + SVM	86.48	87.43	497.156

Table 5. Experimental Results on Farm-ads dataset

Methods	Macro-F (%)	Micro-F (%)	Time (s)
NC	66.38	67.86	80.503
Euclidean KMF + NC	61.75	62.93	0.313
Cosine KMF+ NC	**70.26**	**70.80**	1.313
k-NN(k=5)	81.55	82.24	3695.793
Euclidean KMF + kNN	83.35	83.49	469.766
Cosine KMF + kNN	**84.54**	**84.75**	552.703
SVM	**89.56**	**89.67**	3101.951
Euclidean KMF + SVM	85.67	85.76	661.031
Cosine KMF + SVM	86.45	86.53	640.484

From the results of Table 2 to 5, methods NC and k-NN based on KMF feature selection, all can obviously outperform original NC and k-NN methods, and comparable to SVM in accuracy comparisons. In running time, as dimensionality of data is reduced greatly by KMF, thus the methods by k-means feature are all faster than corresponding original methods.

For text categorization methods based on the k-means feature selection we also discuss the accuracy (Micro-F) with different number of cluster centroids for each class, and with different number of features selected in each centroid. We give the results of two datasets, DBworld dataset and Farm-ads dataset, in Fig. 2 and Fig. 3.

On DBworld dataset (see Fig. 2.), the left of figure (a)(b)(c) show the results with different number of cluster centroids (k=1,2,…,5), and fixed feature number, v=50, and the right of figures show the results with fixed number of cluster centroids (k=3), and various feature number (v=10,20,…,100).

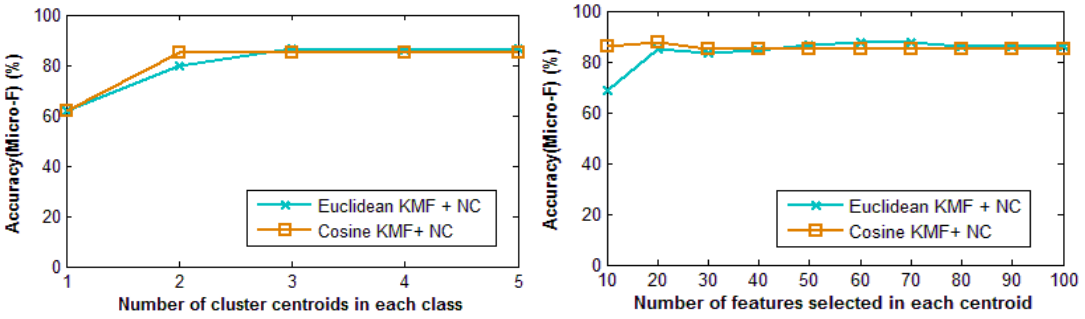

Fig. 2 (a) KMF+NC on DBworld dataset

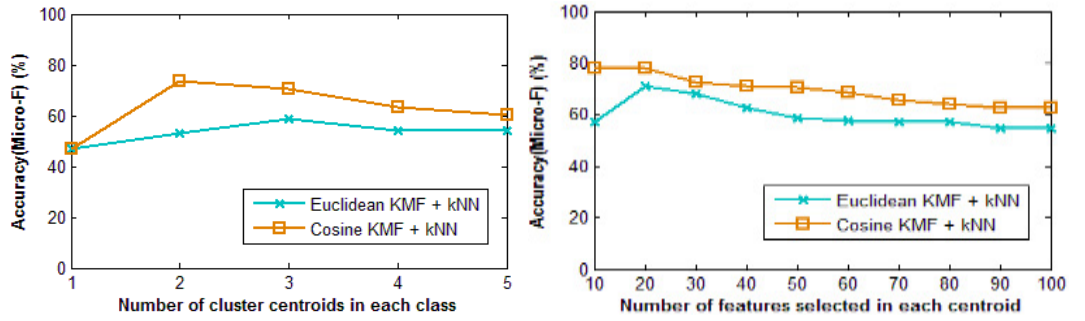

Fig. 2 (b) KMF+kNN on DBworld dataset

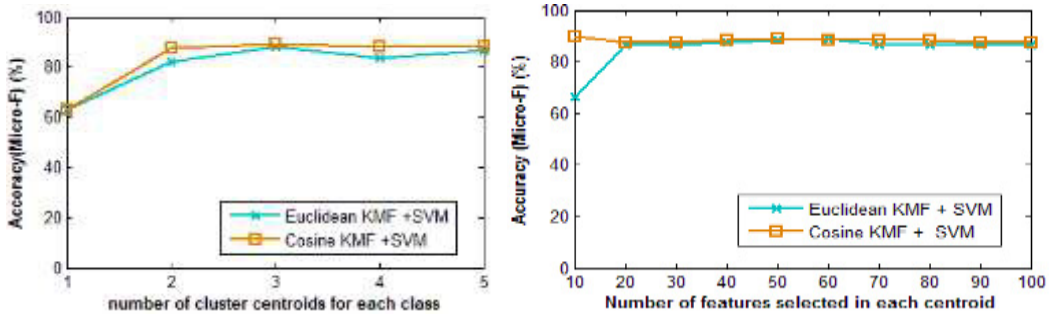

Fig. 2 (c) KMF+SVM on DBworld dataset

Fig. 2. Experimental results of text categorization with KMF on DBworld dataset

On Farm-ads dataset (see Fig. 3.), the left of figure (a)(b)(c) show the results with different number of cluster centroids (k=1,2,…,5), and fixed feature number v=50, and the right of figures show the results with fixed number of cluster centroids (k=4), and various feature number (v=10,20,…,100).

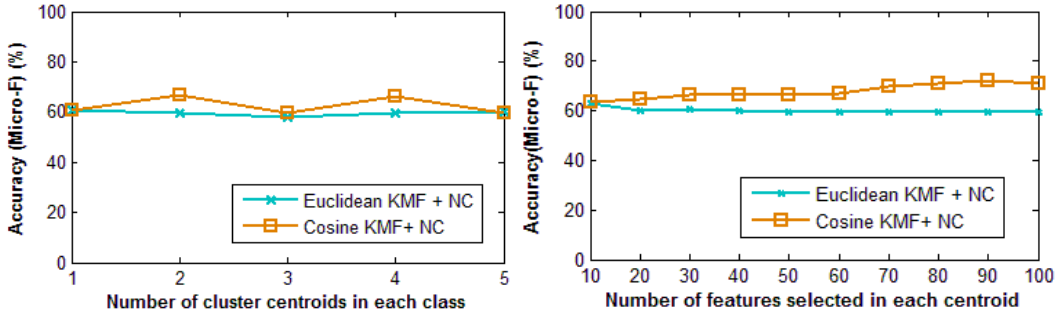

Fig. 3 (a) KMF+NC on Farm-ads dataset

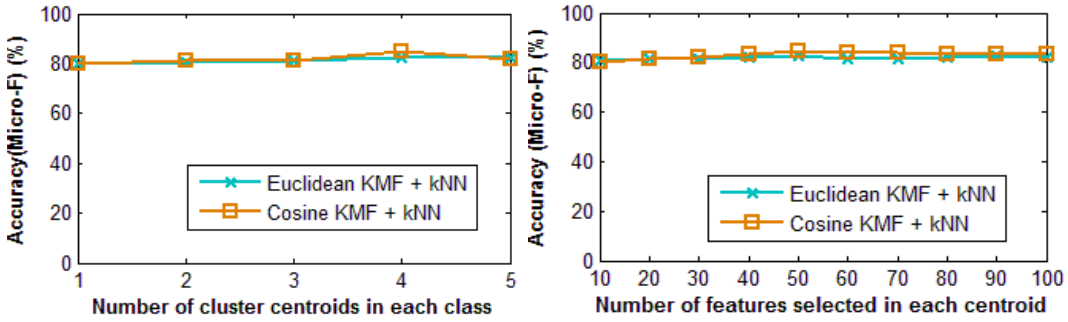

Fig. 3 (b) KMF+kNN on Farm-ads dataset

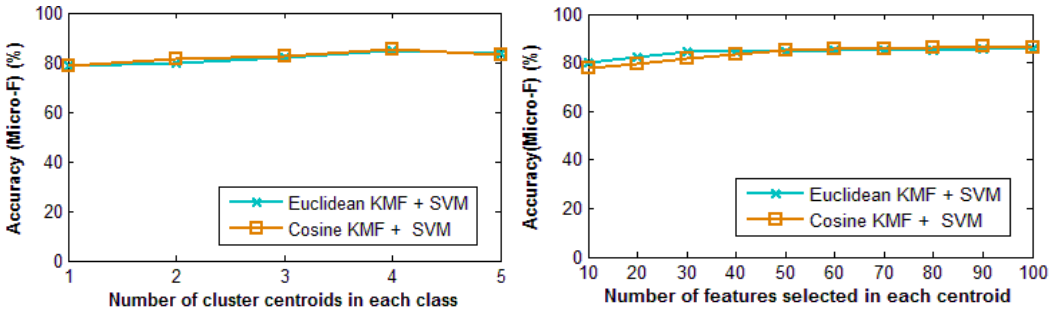

Fig. 3 (c) KMF+SVM on Farm-ads dataset

Fig. 3. Experimental results of text categorization with KMF on Farm-ads dataset

From the results of Fig.2 and 3, we can see that for different datasets, the best choices of centroid number and feature number are different. On DBworld dataset, when the centroid number is 3 and correspondingly the number of chosen feature in each centroid is 20, most of classifiers can reach better results. On Farm-ads dataset, when the centroid number is 4 and the number of chosen feature in each centroid is 50, classifiers basically can reach better results.

Comparing accuracies of three classifiers, SVM is the best one and NC performs better than k-NN on our experiments. For the two similarity distances in k-means, the Cosine distance is more suitable to the text data than Euclidean distance. In all the results in this paper, the methods with Cosine KMF all outperform Euclidean KMF.

4. Conclusions

In this paper, we use k-means clustering method to collect and choose features for text categorization. As the words in clustering centroids of each class can represent class well, thus we choose the features with larger word-frequency for text categorization. Experiments on several text corpus show that the capacities of text classifiers will be enhanced by k-means feature selection.

5. Acknowledgements

This work was supported by Strategic Priority Research Program of Chinese Academy of Sciences (No.XDA06030200), by National Nature Science Foundation of China (No. 61202226).

References

1. Salton, G., McGill, M. editors. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.
2. R. C. Amorim, Learning feature weights for K-Means clustering using the Minkowski metric. University of London. PhD Thesis 2011.
3. Whitney, A. W. A direct method of nonparametric measurement selection. IEEE Transactions on Computers, vol 20, pp.1100-1103. 1971.
4. Dy, J G. (2008) Unsupervised Feature Selection. In: H. Liu and H. Motoda (Ed.) Computational Methods of Feature Selection, Chapman & Hall/CRC, pp. 19-39.
5. Sebastiani, F. Machine learning in automated text categorization. Technical Report IEI-B4-31-1999, Istituto di Elaborazione dell'Informazione, Consiglio Nationale delle Ricerche, Pisa, IT, 1999.
6. Filannino, M. Dbworld e-mail classification using a very small corpus', project of machine learning course, university of manchester.

Technical report, 2011.
7. Agarwal, S., Godbole, S., Punjani, D., and Roy, S. How much noise is too much: A study in automatic text classification. ICDM 2007, 2007.
8. Craven, M., DiPasquo, D., Freitag, D., Mccallum, A., Mitchell, T., Nigam, K., and Slattery, S. Learning to extract symbolic knowledge from the world wide web. AAAI-98, 1998.
9. M. Chris, J. P. Michael. Active learning using on-line algorithms. In KDD 2011, 2011.
10. Ciarelli, P. M. and Oliveira, E. Agglomeration and elimination of terms for dimensionality reduction. In Ninth International Conference on Intelligent Systems Design and Application, pp. 547–552, 2009.